Addendum to:
An Analysis of Information Migration

DMS 98-01-26

The “Analysis of Information Migration” paper took much longer to produce than expected, and in an effort to get it out for consideration at the ISO archiving workshop of Jan. 1998, its summary section is a weak analysis of the migration cases. The purpose of this addendum is to address some of these weaknesses.

The paper continues the Reference Model view of the Content Information being based on a well defined set of bits. The Packaging Information is said delimit these bits, and thus to support access to these bits. The summary section begins to address the reality that, for many current situations, once the Content Information bits have been identified we would often find that the Packaging Information does more than just delimit the bits. It also provides and supports some explicit relationships among some of the Content Information pieces. Some examples can help clarify this situation.

Consider a CD-ROM with a media format that is ISO-9660 compliant. It has been declared that the byte-sequence content of each file, the file names, and the directory hierarchy, compose the Content Information. (Whether this really makes sense is a separate issue, but it probably does for some situations.) We can envision creating distinct data objects for each of these pieces of information, and documenting their representations to form ‘complete content information’. However, there also would need to be an over-arching representation that described the relationships among them, including how each entry in the directory hierarchy was related to file names and contents. When this information is in the context of an ISO-9660 media format, which is the Packaging Information, this over-arching representation is actually provided by the Packaging Information. This Packaging Information plays a much bigger role than simply delimitation - it also specifies specific relationships.

Two avenues for migration to new media types (such as a DVD) suggest themselves in this example. The first is to find new Packaging Information that supports not only the same delimitations, but also the same relationships among the delimited objects. I believe this would be a case of what Randy called ‘congruence’ in the current Reference Model migration section 5 - i.e., there is a one-to-one mapping between the delimited objects before and after the Repackaging. However I believe more is at stake here, because there also needs to be the same semantic relationships among the delimited objects. Because DVDs support the same data model as CD-ROMs, at the level we are requesting, one could move the Content Information of this example into a DVD in the natural way and the new Packaging Information would do the job. From this example it seems clear that when the Packaging Information is relied upon to support specific relationships, the ability to easily migrate to new media types is limited.

The second avenue for migration, in the example above, is to attempt to remove the ‘specific relationship’ requirements from the Packaging Information by incorporating the specific relationship information into the Content Information itself. This could be done by creating another data object, which contained the relationship information, and the ability to point to these other objects in a generic (i.e., easily migrateable) way. The question then arises as to what are the simplest, and yet practical, requirements for Packaging Information in such a context? Note that Packaging Information that conformed to such requirements would not be limited to only this level of support, but this level of support would be all that was needed to preserve Content Information during migrations. Of course the Content Information would need to include all its specific relationships.

If the Packaging Information was relied upon to only delimit the pieces and the whole, this implies that it supports a way to identify each piece and a way to identify that each piece is a part of a whole. It should also identify that the whole is made up of these pieces. Then any further relationships among the pieces might be the responsibility of the Content Information to specify. The Content Information would have to be able to refer to the pieces, and possibly to sub-pieces, in a non-ambiguous and migrateable way. One approach to implementing the reference technique would be to use the same handles that the Packaging Information used to identify the pieces. This could be constraining unless the same types of handles could be used in a variety of packages. Another approach would be to use handles embedded in the Content Information, or some combination of the two.

Considerably more analysis is needed to flesh this out, but it seems that some standard packaging approaches could be adopted that would facilitate migration across various media types without having to update the Content Information. If this were the case, the argument could be made that such migrations do not need to have Provenance Information updated. They are no more disruptive than moving microfilm to a new storage bin, for example. This appears to be a desirable objective for migrations in an archive.

An issue that is not addressed is the potential impacts of fixity information in migration. This clearly relates to Packaging Information, and may have an implementation that looks like another object packaged with the Content Information. However some examples need to be given.

I believe another issue that is not adequately addressed is the desire for certain types of Packing Information to support Consumer preferences. If the archive chooses to go with generic, easily migrateable Packaging Information objects, then it may need to remap these to Consumer desired packages upon request. However the very fact of having the Content Information within standard packages facilitates automated mapping to a number of distributable package types.

Yet another issue that could use more discussion is the realities associated with Transmutations. The paper talks about some of these issues. Randy has identified the principal of equivalence, which obtains when there is a reverse transformation that exactly reproduces the original from the Transmutated form. This implies that information has not been lost as long as the transformation has been documented and is 1-to-1 reversible. I believe this is the same effect as requiring that the corresponding representations have overlapping sets of semantically equivalent ‘resultant entities’ and all these ‘resultant entities' are unique.   This certainly works for ASCII to UNICODE and back, and also for 'column major' to 'row major' and back. With regard to updating Provenance when such Transmutations are made, it appears that all Transmutations to Content Information should be documented in Provenance. However, what should we say about updates to Context, Reference, or Fixity (a new technique is used, for example)? Perhaps the archive should keep an update log for these object types? Where - Archival Storage or just Data Management?

=====